Segmentation temporelle et spatiale de données agricoles

Jean-François Mari, Florence Le Ber, Marc Benoît

Segmentation temporelle et spatiale de données agricoles

Jean-François Mari * — **Florence Le Ber** * — **Marc Benoît** **

* *LORIA : UMR 7503 et Inria-Lorraine, B.P. 239, 54506 Vandœuvre-lès-Nancy*
jfmari@loria.fr, leber@loria.fr
** *INRA SAD, Domaine du Joly, 88400 Mirecourt*
benoit@mirecourt.inra.fr

RÉSUMÉ. *Nous présentons les résultats d'un travail de fouille de données effectué par des agronomes et des informaticiens pour extraire des bases de données agricoles Ter-Uti des informations sur les successions de cultures pratiquées dans une région. Nous avons utilisé pour cela des modèles stochastiques développés pour l'analyse de séquences et d'images : les modèles de Markov cachés. Ces modèles permettent de représenter des observations temporelles et spatiales comme des successions d'états où les transitions entre états dépendent, suivant l'ordre du modèle, de l'état courant et des n états voisins. Nous présentons une méthode de classification spatio-temporelle qui permet de découper une région sur la base des évolutions des successions de cultures qui y sont pratiquées en faisant apparaître des classes homogènes de successions de cultures. Deux exemples illustrent notre travail : les cas de la Lorraine et de Midi-Pyrénées.*

ABSTRACT. *In the area of data mining of agricultural databases, this paper presents some results in segmentation and classification of spatio-temporal data using high-order Hidden Markov Models (HMM). These models are capable to map spatial and temporal observations onto a sequence of states in which the transitions between the states depend on the n previous states according to the order of the Markov chain. They have been applied on spatial and temporal data concerning land use, named Ter-Uti data, in order to find agricultural land use regularities. We show on various examples that the HMMs are powerful tools for temporal and spatial data mining.*

MOTS-CLÉS : *HMM, données Ter-Uti, utilisation du territoire, fouille de données, segmentation spatio-temporelle.*

KEYWORDS: *HMM, Ter-Uti data, data mining, land use, spatio-temporal segmentation.*

Actes des $6^{èmes}$ Journées CASSINI 2002, pages 251 à 272

Introduction

L'occupation du territoire agricole change d'années en années, en raison de deux faits majeurs, la libération continue de territoires par la disparition d'exploitations agricoles et l'évolution des systèmes de production. Cette évolution se traduit principalement dans le changement des successions de cultures, par l'introduction de nouvelles cultures, le raccourcissement des rotations, la prise en compte de la réglementation de la politique agricole commune, etc.

L'étude de ces successions est un enjeu pour l'agronomie car :

– la connaissance des successions et de leur évolution est un outil pour la gestion de certaines ressources de l'agriculture (par exemple l'irrigation) ;

– c'est également un moyen de prévoir et de prévenir des effets environnementaux majeurs tels que la pollution des ressources en eau et la structuration des paysages ;

– à un niveau plus prospectif, cette connaissance est utile pour analyser les dynamiques en cours dans l'agriculture.

Le choix d'une succession par un agriculteur reste un phénomène mal connu. Ce choix est révélateur du métier d'agriculteur et intègre de nombreux facteurs : dates de récolte et d'implantation des cultures, état laissé par une culture après la récolte, organisation de chantiers entre parcelles, etc.

Différents moyens existent cependant pour connaître les successions pratiquées dans une région : enquêtes de terrain chez les agriculteurs, recueils d'expertise auprès de techniciens agricoles ou exploitation des bases de données, constituées en particulier par les services statistiques du ministère de l'agriculture. C'est ce dernier moyen que nous avons étudié. Nous avons en effet développé une méthode pour exploiter la base de données *Ter-Uti*, qui est un recueil systématique depuis 20 ans des occupations du sol sur un ensemble de points du territoire français [LED 92]. Notre but est d'une part de définir des régions homogènes pour les successions culturales (segmentation spatiale), d'autre part de mettre à jour les successions dominantes et de cerner leurs évolutions à différentes échelles (segmentation temporelle).

Cet article présente la méthode que nous avons développée conjointement et qui s'inscrit dans un processus de fouille de données entre informaticiens "fouilleurs" et agronomes "experts" [MAR 00]. Nous avons utilisé des algorithmes précédemment développés pour la reconnaissance de la parole que nous avons adaptés aux données *Ter-Uti*. Nous avons développé des outils pour manipuler les données et visualiser les résultats en fonction des demandes des agronomes qui les ont ensuite pris en main et utilisés pour leur recherche. Des résultats sont présentés sur la région Lorraine, qui a servi de région test, et sur la région Midi-Pyrénées où notre méthode est utilisée dans le cadre d'un projet d'aide à la gestion de l'irrigation.

Le plan de l'article est le suivant : la partie 1 présente le principe de la méthode et les modèles théoriques utilisés, les modèles de Markov cachés. La partie 2 présente les données *Ter-Uti*. La partie 3 décrit les outils mis au point pour la segmentation

temporelle et les résultats obtenus sur les données lorraines. La partie 4 décrit la méthode utilisée pour la segmentation spatiale et les résultats obtenus en Lorraine et en Midi-Pyrénées. La dernière partie est une discussion.

1. La méthode utilisée : fouille de données par modèles de Markov cachés

Une des fonctions de la fouille de données est d'élaborer des indices concis et expurgés du bruit de l'application qui permettent à un analyste d'avoir une vue claire d'un tableau de nombres. A partir de cette vue, l'analyste peut extraire quelques éléments de connaissance, débuter un raisonnement et formuler d'autres expérimentations susceptibles de compléter sa connaissance du domaine.

La classification reste une des approches principales en fouille de données. Dans cette optique, nous adoptons une approche bayésienne. Afin de spécifier et construire des interfaces de visualisation de résultats, nous utilisons des indices élaborés à partir de la probabilité *a posteriori* d'effectuer un classement une fois les données observées.

Les modèles de Markov cachés (HMM comme *Hidden Markov Model*) permettent de calculer cette probabilité dans le cas où les données ont une dimension temporelle et/ou spatiale. Ils seront présentés quelques lignes plus bas dans la section 1.2.

Nous allons envisager dans la suite deux familles de tableaux de données numériques qu'il faudra classer à l'aide de HMM : les suites temporelles d'observations comme les cultures relevées dans un point précis – un site – du territoire au fil des ans et les champs de Markov fondés sur la notion de voisinage spatial comme la répartition spatiale des cultures sur un territoire une année donnée. Nous verrons comment les hypothèses markoviennes permettent de traiter dans un cadre mathématique élégant la variabilité temporelle et spatiale des formes étudiées et comment la variabilité temporelle d'un site peut influencer sa catégorisation spatiale.

L'analyse des changements d'occupation des sols est un thème de recherche actuel en agronomie. Nous l'abordons avec une méthode de classification qui a fait ses preuves en reconnaissance de la parole et en génétique. La poursuite de ce travail consistera en une évaluation complète ainsi qu'une comparaison de cette approche aux approches utilisées en analyse des changements d'occupation des sols [KIE 93].

1.1. *Modèles de Markov cachés pour la classification*

Dans notre domaine, les sites ont des caractéristiques propres – l'occupation de la parcelle contenant ce site – qui permettent de les regrouper en zones homogènes mais aussi s'inscrivent dans une évolution temporelle qui peut constituer un critère de classification supplémentaire. Le classement d'un point dépend autant de ses caractéristiques aux instants t, $t-1 \ldots$ que de celles de ses voisins spatiaux. Pour simplifier la modélisation, nous acceptons l'hypothèse markovienne qui stipule que, sur une parcelle, la règle de succession ne dépend que de l'occupation actuelle de la parcelle et

de l'occupation des deux années précédentes ainsi que des occupations des points voisins. Ces hypothèses nous permettent d'utiliser les modèles de Markov cachés d'ordre un et deux (HMM1 et HMM2). Les experts agronomes acceptent facilement ces hypothèses.

1.2. *Définition d'un HMM*

Les modèles de Markov cachés dérivent des chaînes de Markov. Une chaîne de Markov définit un seul processus stochastique : elle possède un ensemble d'états – par exemple les cultures successives d'une parcelle – directement observables. Dans un HMM, une observation n'est pas uniquement associée à un état. Les états sont définis comme des densités de probabilité sur l'ensemble des observations. Dans notre application, on considère ainsi que la répartition – *c.-à-d.* la densité – des cultures dans une région donnée évolue selon un processus de Markov. La répartition à un pas de temps donné ne dépend que de la répartition aux pas précédents suivant l'ordre du modèle. On suppose ainsi que l'observation d'une culture une année donnée sur une parcelle suit la loi définie sur l'état atteint cette année par la chaîne de Markov.

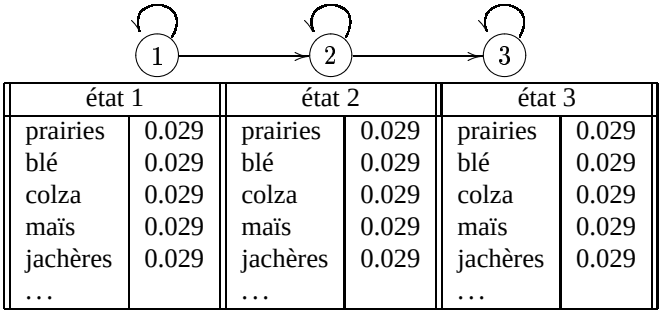


Figure 1. *HMM effectuant une segmentation en trois périodes. Les états cachés sont dénotés 1, 2 et 3*

Le HMM permet de représenter deux processus stochastiques, le premier gouvernant le second [BAK 74] :

– le premier processus est défini sur un ensemble d'états et caché pour un observateur. C'est une chaîne de Markov d'ordre un ou deux.

– le deuxième processus est qualifié de visible. Il émet une observation à chaque pas de temps en fonction des densités de probabilité définies sur chacun des états par le processus caché.

Dans l'exemple donné Fig. 1, les états cachés sont dénotés 1, 2 et 3. Ils représentent chacun une distribution uniforme de cultures. Ce modèle peut simuler la production d'une région. A chaque pas de temps, la chaîne de Markov change d'états en fonction des transitions autorisées et produit une occupation du sol : prairies, blé, colza, jachères, ... en fonction de la densité représentant l'état visité.

Finalement, un modèle de Markov caché est défini par la donnée de :

– $\mathcal{S} = \{O, s_1, s_2, \ldots, s_N, F\}$, un ensemble fini comprenant $N + 2$ états, dont un état initial O et un état final F ;

– $\mathcal{A}$ la matrice donnant les probabilités de transition entre états :

- $\mathcal{A} = (a_{ij})$ pour un modèle d'ordre 1 (HMM1) avec la contrainte :

$$\sum_j a_{ij} = 1 \quad \forall\, i$$

- $\mathcal{A} = (a_{ijk})$ pour un modèle d'ordre 2 (HMM2), avec la contrainte :

$$\sum_k a_{ijk} = 1 \quad \forall\, i, j$$

– $b_i(.)$ les lois des densités associées aux états s_i.

La matrice $\mathcal{A}$ est initialisée avec un ensemble de valeurs qui permettent de définir la topologie du graphe des transitions entre états : quelles sont les transitions autorisées, aller simple ($a_{ij} > 0$, $a_{ji} = 0$), aller-retour ($a_{ij} > 0,\ a_{ji} > 0$), bouclage ($a_{ii} \neq 0$), etc. Par exemple, le HMM1 décrit par la figure 1 possède pour matrice de transitions :

$$\mathcal{A} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}$$

La modélisation d'un signal temporel par des HMM est ainsi fondée sur deux principes : (i) le signal temporel peut être découpé en segments par une chaîne de Markov et (ii) le signal est la réalisation d'un processus stationnaire représenté par une densité de probabilité sur l'espace des observations à l'intérieur d'un segment.

En modélisant de la sorte les segments composant le signal temporel, on ignore la réalité de la production du signal comme résultat d'un processus intelligent mais on peut utiliser des algorithmes d'apprentissage et de reconnaissance rapides [BOY 90]. Cette façon de procéder est complémentaire d'une approche analytique et explicative fondée sur un mécanisme de raisonnement. En mesurant précisément par une probabilité ce que l'on qualifie au premier abord de hasardeux, on diminue l'indéterminisme de notre perception du processus et on peut faire apparaître des comportements explicables donc prévisibles qui pourront être réutilisés dans un mécanisme de raisonnement ; ce mécanisme d'extraction et de réutilisation est un des principes de la fouille de données.

1.3. *Probabilité d'un alignement temporel*

En adoptant les notations définies par Rabiner dans [RAB 95], on appelle Q une fonction d'un intervalle de temps fini $[1, T]$ dans l'ensemble $\mathcal{S}$ des états.

$$Q \ : \ [1, T] \rightarrow \mathcal{S}$$

On définit un alignement temporel entre une suite d'états $Q = q_1, q_2, \ldots, q_t, \ldots q_T$ atteints respectivement aux temps $1, 2, \ldots, t, \ldots T$ et une suite de observations $O = o_1, o_2, \ldots, o_t, \ldots o_T$ par une correspondance (état, observation). Sa probabilité est définie par :

$$Prob(Q, O) = \prod_t a_{q_{t-2}q_{t-1}q_t} b_{q_t}(o_t) \quad (1)$$

où :

– $a_{q_{t-2}q_{t-1}q_t}$ représente la probabilité de la contribution de la chaîne de Markov entre les instants $t-2$ et t (cas d'une chaîne d'ordre 2) qui définit la segmentation ;

– $b_{q_t}(o_t)$ représente la vraisemblance de l'observation o_t conditionnée à la densité de l'état q_t.

Dans le cas des modèles de Markov gauche-droite décrits dans la figure 1, un alignement entre états et observations définit une segmentation.

La vraisemblance $L(O)$ est calculée en considérant tous les alignements possible.

$$L(O) = \sum_Q Prob(Q, O) \quad (2)$$

La représentation d'un phénomène temporel ou spatial à l'aide d'un HMM peut avoir plusieurs objectifs :

– **estimation** ou apprentissage de paramètres. Par exemple on veut calculer les probabilités des successions de cultures à la lumière des observations faites dans une région.

– **discrimination** ou reconnaissance. Par exemple, une fois le modèle estimé, on veut retrouver la suite d'états la plus probable qui explique une suite d'observations.

– **segmentation** par exemple, on recherche une date de changement entre deux états donnés.

Un grand intérêt des HMM est qu'ils possèdent des algorithmes polynomiaux dans chacun de ces trois domaines. Ces algorithmes sont fondés sur la recherche de tous les alignements afin de retrouver celui possédant la probabilité définie par l'équation 1 maximale [FOR 73, BAK 74].

1.4. *Estimation automatique d'un HMM*

Une fois donné un corpus de données et la topologie du graphe des transitions entre états, différents algorithmes permettent l'apprentissage d'un HMM et la détermination des paramètres $\mathcal{A}$ et $b_i(.)$ (cf.§1.2). Quel que soit l'ordre des modèles, nous utilisons l'algorithme Forward - Backward [MAR 97] qui est une variante de l'algorithme EM [DEM 77]. L'apprentissage se fait itérativement en partant d'un modèle où toutes les

transitions sont équiprobables et où les densités $b_i(.)$ sont fixées. L'algorithme Forward - Backward calcule un nouveau modèle plus adapté aux données dans lequel la vraisemblance du corpus a augmenté. Ce nouveau modèle est utilisé dans une nouvelle itération jusqu'à ce que la vraisemblance du corpus atteigne un maximum local. Le résultat est constitué par les nouvelles valeurs des transitions a_{ijk} et des densités $b_i(.)$ (cf. Fig 2). Lorsque N est le nombre d'états et T le nombre d'observations, l'algorithme Forward - Backward a une complexité en N^3T pour un HMM2.

L'estimation des paramètres d'un HMM2 se fait grâce au calcul des probabilités suivantes.

La fonction $\alpha_t(j,k)$ definit la probabilité de la séquence partielle d'observations, $o_1, ..., o_t$ et de la transition (s_j, s_k) entre les instants $t-1$ et t :

$$\alpha_t(j,k) = \mathrm{P}(o_1, o_2, ..., o_t, q_{t-1} = s_j, q_t = s_k), \quad \begin{array}{ll} 2 & \leq t \leq T, \\ 1 & \leq j,k \leq N. \end{array} \tag{3}$$

$\alpha_t(j,k)$ est calculée par récurrence à partir de $\alpha_{t-1}(i,j)$ dans lequel (s_i, s_j) et (s_j, s_k) sont deux transitions entre les états s_i et s_k :

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{N} \alpha_t(i,j).a_{ijk}.b_k(o_{t+1}), \quad \begin{array}{ll} 2 & \leq t \leq T-1, \\ 1 & \leq j,k \leq N. \end{array} \tag{4}$$

D'une façon similaire, la fonction $\beta_t(i,j)$ définit la probabilité de la séquence partielle d'observations de $t+1$ à T en supposant la transition (s_i, s_j) entre les instants $t-1$ et t :

$$\beta_t(i,j) = \mathrm{P}(o_{t+1}, ...o_T | q_{t-1} = s_i, q_t = s_j), \quad \begin{array}{ll} 2 & \leq t \leq T-1, \\ 1 & \leq i,j \leq N. \end{array} \tag{5}$$

Cette fonction se calcule récursivement :

$$\beta_t(i,j) = \sum_{i=1}^{N} \beta_{t+1}(j,k).a_{ijk}.b_k(o_{t+1}), \quad \begin{array}{ll} 2 & \leq t \leq T-1, \\ 1 & \leq j,k \leq N. \end{array} \tag{6}$$

Une fois donnée la séquence d'observations O, nous définissons $\eta_t(i,j,k)$ comme la probabilité *a posteriori* de la transition $s_i \longrightarrow s_j \longrightarrow s_k$ entre $t-1$ et $t+1$ pendant l'émission de O :

$$\eta_t(i,j,k) = \mathrm{P}(q_{t-1} = s_i, q_t = s_j, q_{t+1} = s_k / O), \quad 2 \leq t \leq T-1.$$

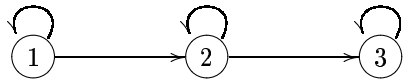
On déduit :

$$\eta_t(i,j,k) = \frac{\alpha_t(i,j) a_{ijk} b_k(o_{t+1}) \beta_{t+1}(j,k)}{L(O)}, \quad 2 \leq t \leq T-1. \tag{7}$$

Cette quantité permet le calcul de la nouvelle estimée de a_{ijk}.

$$a_{ijk} = \sum_t \eta_t(i,j,k). \tag{8}$$

Le choix du modèle initial influe sur le résultat final. Pour évaluer l'adéquation du modèle obtenu par convergence, nous utilisons une mesure de distance entre états [TOU 74]. Par essais successifs, nous définissons et obtenons un modèle dépendant de plusieurs facteurs : résultats des expériences précédentes, topologie initiale, mode de convergence, critère d'arrêt.



état 1		état 2		état 3	
prairies	0.31	prairies	0.29	blé	0.29
blé	0.22	blé	0.26	prairies	0.27
orge	0.16	colza	0.14	colza	0.17
colza	0.12	orge	0.11	orge	0.12
maïs	0.07	maïs	0.08	maïs	0.06
jachères	0.05	jachères	0.05	vergers	0.02

Figure 2. *Modèle 1 : HMM de la figure 1 après apprentissage sur les données issues d'une région agricole. Les distributions de cultures obtenues sont caractéristiques de la région*

2. Un exemple de données spatio-temporelles agricoles : les données *Ter-Uti*

Notre ensemble de données est constitué de l'enquête *Ter-Uti* qui est réalisée par un sondage à deux niveaux de granularité. Un premier tirage, réalisé par l'IGN, consiste à sélectionner des photos aériennes régulièrement réparties sur l'ensemble du territoire métropolitain. Les photos représentent chacune un carré de 2 km de côté et sont séparées en moyenne par 6 km. Un deuxième tirage, réalisé par les DRAF[1], consiste à superposer sur chaque photo, une grille de 36 points. Compte tenu de la distance entre les photos, la représentativité d'un point est proche de 100 hectares. L'ensemble de ces sites est visité annuellement par des enquêteurs qui relèvent les occupations des sites. Pour plus de détails sur la grille *Ter-Uti*, on peut se reporter à [LED 92].

Nous disposons de résultats sur différentes régions – Lorraine (23756 points) et Midi-Pyrénées (10904 points) – pour les années 1992 à 2000. Outre la séquence temporelle des occupations de chaque point, nous savons à quelle PRA[2] il appartient et

1. Direction Régionale de l'Agriculture et de la Forêt.
2. PRA pour Petite Région Agricole.

nous connaissons ses voisins, c'est-à-dire la disposition relative de chaque point et de chaque photo aérienne. En revanche, nous ignorons la localisation précise des points pour des raisons de secret statistique.

Les services de statistique de la DRAF ont réparti les occupations en différentes classes (environ 80) qui vont de « marais salants, étangs d'eau saumâtre » à « peupliers épars » en passant par « superficie en herbe à faible productivité potentielle ». Certaines de ces classes ne sont pas ou peu présentes dans les régions étudiées considérées aussi avons nous restreint le nombre de classes à 49, par regroupement ou suppression [BEN 01].

3. Classification temporelle des données *Ter-Uti* lorraines

3.1. *Étude des répartitions de cultures dans le temps*

Nous nous sommes tout d'abord intéressés à la définition de périodes d'observation pendant lesquelles la distribution des cultures ne varie pas. Pour la période étudiée (1992-99) nous nous limitons donc à des modèles possédant deux ou trois états mais autorisant des transitions en boucle sur les états comme le montre la figure 1. Ce modèle permet de segmenter la période étudiée en trois sous-périodes d'environ trois ans où le système est supposé stationnaire. On justifie *a posteriori* le découpage en trois états sur 8 ans par le fait que trois ans correspondent à l'échelle de temps d'une rotation en Lorraine.

La figure 2 illustre un résultat de classification dans laquelle nous avons cherché trois situations stables de production agricole. Entre les années 92 et 99, la région étudiée (PRA 306[3]) est passée par trois états de distributions de cultures différentes. Dans ces distributions, on voit immédiatement l'importance des prairies (de l'ordre de 30 %), la progression du blé et du colza au détriment de l'orge, et la disparition de la jachère au fil des ans. Sur la PRA 306 et d'autres régions et à partir de ces résultats, nous avons déterminé les occupations à examiner prioritairement dans l'ensemble des occupations de la base *Ter-Uti* ; c'est-à-dire à la fois les occupations les plus fréquentes et les plus instables *a priori* : blé, orge, maïs, colza et prairies pour la Lorraine.

3.2. *Étude de successions de cultures*

Dans un HMM tel que défini dans la figure 1, il n'est pas possible de mesurer la probabilité d'une succession de cultures puisqu'une culture n'apparaît qu'à l'intérieur d'une répartition constituant l'état. Pour pallier ce défaut, nous introduisons des états uniquement associés aux cultures majoritaires que nous voulons étudier (blé, maïs, orge, ...). Le HMM obtenu possède deux types d'états : les états "de réserve" qui sont associés normalement à des répartitions de cultures – comme dans le modèle 1 – et

3. La PRA 306 est le plateau lorrain sud.

les états “de Dirac” qui sont associés à une culture seule et définis par une densité où la probabilité de cette culture vaut 1 et les probabilités de toutes les autres occupations valent 0. Les états de réserve permettent de faire une segmentation progressive des données. Les cultures qui apparaissent comme étant les plus fréquentes dans un état de réserve, ou qui nous intéressent particulièrement, peuvent être placées dans un nouvel état de Dirac dans une expérience suivante. Finalement la figure 3 donne la topologie d'un modèle 2 qui sera utilisé par la suite. Ce modèle permet d'étudier les transitions entre les états associés au blé, orge, maïs, colza et prairie. Les états cachés sont dénotés 2, 3 et 4 et jouent le rôle d'états de réserve.

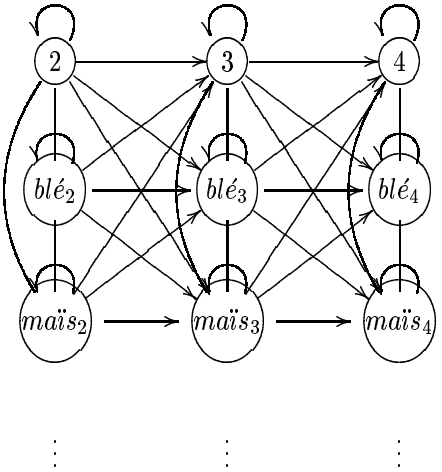


Figure 3. *Modèle 2 : les états notés 2, 3 et 4 sont associés à une distribution de cultures, contrairement aux états dénommés par une occupation. Le nombre de colonnes définit le nombre de périodes d'observation. Les connexions sans flèches représentent des transitions bidirectionnelles*

Les résultats obtenus par le modèle 2 sont visualisés au moyen d'un graphe tel que présenté en figure 4 pour la PRA 306. Les transitions entre états de Dirac calculées par l'équation 7 sont représentées par des lignes brisées dès que leur probabilité atteint un certain seuil. L'épaisseur du trait est proportionnelle à la valeur de la probabilité. L'agronome en fait l'interprétation suivante.

“Nous voyons sur ce graphique deux ensembles de couverts pérennes qui n'échangent aucune surface avec les autres couverts sur toute la période : les prairies permanentes productives (notées *ppp*), les bois ; de plus apparaissent deux petits groupes de monocultures : la monoculture de blé qui reste stable sur la période, et la monoculture de maïs qui diminue en fin de période.

L'orge semble changer de statut dans les successions culturales : d'un rôle de “tampon” à celui d'élément stable d'une rotation :

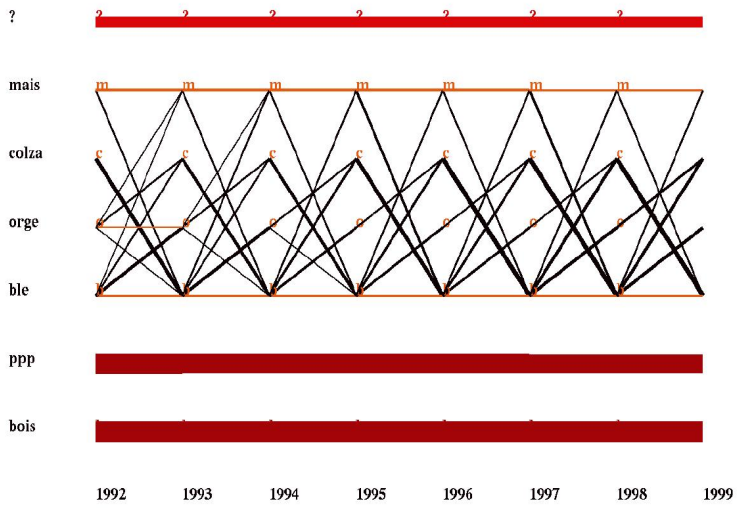


Figure 4. *Résultats du modèle 2 sur les données de la PRA Plateau lorrain sud (PRA 306) entre 1992 et 1999. La ligne dénotée " ?" correspond à un état de réserve affiché et celle dénotée par* ppp *correspond aux prairies permanentes productives*

– au début de la période, l'orge précède un colza ou un maïs ou même un blé et existe en couple orge-orge,

– à la fin de la période, se lit une forte régularité blé-orge-colza, où l'orge n'est plus suivie que de colza. L'orge permet en effet une préparation précoce du sol pour le colza car sa récolte est tôt en saison.

A contrario, le blé reste à une place semblable dans la succession : il est précédent à maïs, colza, orge . . . ou à lui-même".

3.3. *Étude des évolutions de successions*

Nous avons ensuite tenté de déterminer les successions majoritaires en fixant leur taille à 3. Cette valeur n'est pas arbitraire mais correspond à une observation sur la Lorraine : dans la presque totalité des cas, les successions s'organisent en rotation avec des têtes de rotation qui reviennent au plus tous les trois ans. Tous les triplets possibles de la base sont considérés (6 triplets pour 8 années, donc 81915 pour 23756 points, mais seulement 1109 triplets différents). Chaque triplet constitue une observation de trois occupations successives faites pendant la période 1992-99. Les résultats de la classification ont été étudiés sous forme de simples tableaux et ont permis de :

– repérer les successions dans les triplets, c'est-à-dire regrouper les différentes permutations : on vérifie ainsi que les triplets (colza blé orge), (blé orge colza) et (orge

colza blé) ont à peu près la même représentation dans chacun des états ; de même pour (blé colza blé) et (colza blé colza) ;

– repérer et évaluer les successions majoritaires : on vérifie que deux successions (colza blé orge) et (colza blé) représentent une grosse partie des terres cultivées (environ 28 %) ;

– étudier la progression, l'apparition ou la disparition des différentes successions dans la période considérée.

À l'issue de cette analyse, nous avons défini les successions à étudier davantage sous forme de grandes classes ou de rotations types (colza+2céréales, colza+1céréale, maïs+2céréales, maïs+1céréale, monocultures). Nous avons alors, comme lors de la première étape, construit un nouveau modèle – appelé modèle 3 – dans lequel des états n'émettent que ces triplets et leurs permutations circulaires (cf. figure 5).

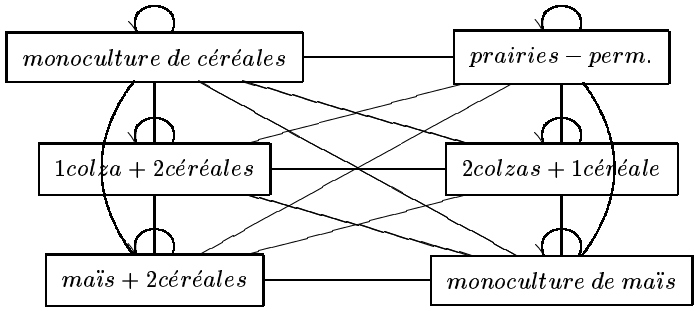


Figure 5. *Modèle 3 : modèle de triplets, similaire au modèle 2. Un état particularise une rotation type. Toutes les transitions sont bidirectionnelles*

Les résultats de ce modèle peuvent être présentés de la même façon que sur la figure 4. Comme une observation est un triplet correspondant à trois observations *Ter-Uti* aux années n, $n+1$ et $n+2$ du même point, on l'indice par son année de début. Ainsi le dernier triplet (1997-1998-1999) est noté 1997. Les transitions entre rotations sont indiquées par les lignes obliques, les variations de l'importance d'une même rotation sont représentées par l'épaisseur du trait horizontal, dès que la probabilité *a posteriori* dépasse 1,0 %. Cette valeur correspond à un seuil d'affichage ajusté par l'utilisateur. En effet, l'affichage des résultats peut être modifié par l'utilisateur, et donc être plus ou moins lisible, et faire apparaître plus ou moins d'informations.

Les trajectoires entre états mises en évidence dans la figure 4 deviennent des lignes droites dans les figures 6 et 7. Ces deux figures représentent les résultats obtenus sur la même région (PRA 306), avec deux seuils différents : sur la figure 6 ce seuil est fixé à 1 % et à 0,4 % sur la figure 7.

Ainsi, l'agronome va utiliser différemment les deux figures 6 et 7. Il a beaucoup de mal de travailler au seuil de 0,4 %. En revanche, avec le seuillage à 1,0 %, il trouve des tendances très instructives :

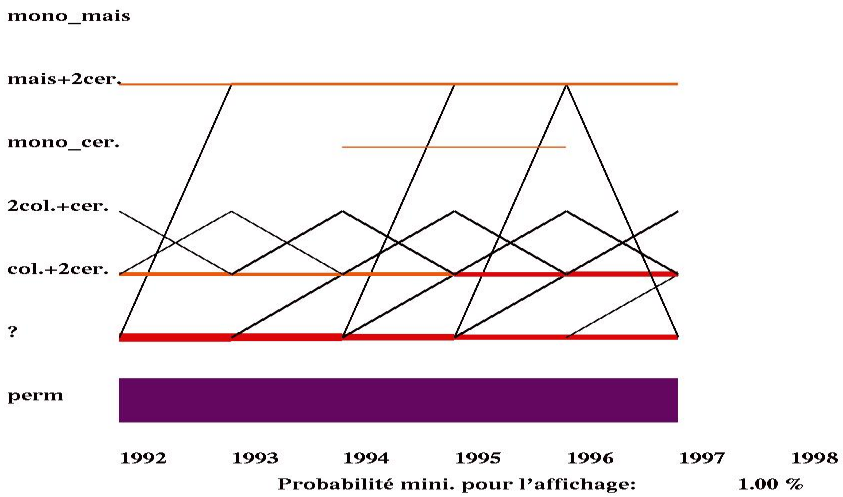


Figure 6. *Résultats obtenus à partir du modèle 3 sur les données de la PRA Plateau lorrain sud entre 1992 et 1999*

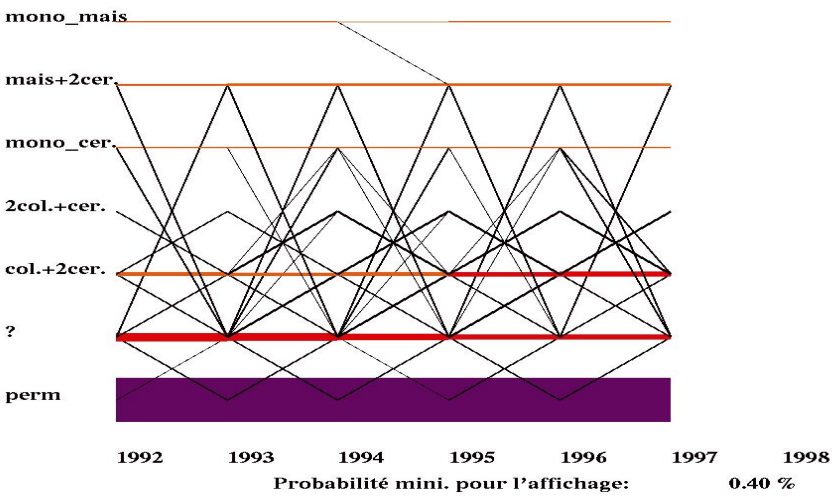


Figure 7. *Résultats du modèle 3 sur les données de la PRA Plateau lorrain sud entre 1992 et 1999*

– “la classe des monocultures de prairies domine largement toutes les autres et reste stable (aucune évolution de type retournement de prairies ou remise en herbe de culture n’apparaît) ;

– l’état de réserve diminue pour alimenter les successions colza + 2 céréales et maïs + 2 céréales ;

– les transitions sont nombreuses entre colza + céréale et colza + 2 céréales, ce qui signifie que sur une même parcelle ces deux successions peuvent avoir lieu (de plus en plus vrai plus on arrive en fin de période) ;

– les monocultures de céréales sont peu nombreuses, instables dans le temps, et non reliées à d’autres successions ;

– la succession maïs + 2 céréales est très stable et augmente légèrement en fréquence au cours de la période.

En conclusion, cette région reste à dominance de monoculture prairiale, avec une tendance à accroître les successions (colza ou maïs) + 2 céréales dans les parcelles cultivées.”

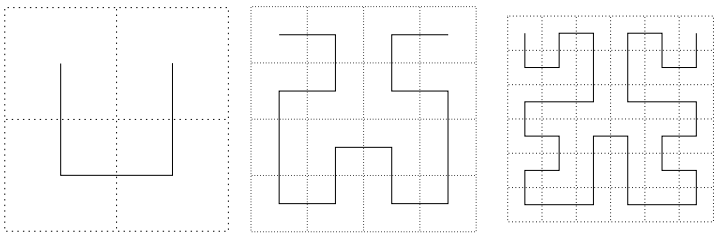
Une étude complète a été menée sur les PRA lorraines entre les années 92 – 2000. Des résultats complémentaires sont donnés dans [BEN 01].

4. Classification spatiale

Dans un espace de dimension deux, comme une image, la notion de chaîne de Markov se généralise en donnant naissance à la notion de champ de Markov : l’état d’un point – appelé aussi un site – dépend seulement de ses plus proches voisins spatiaux. Les algorithmes d’estimation et de classement deviennent alors plus complexes à mettre en œuvre. Au contraire, en suivant l’exemple de [BEN 95], nous avons introduit une relation d’ordre sur les points du plan qui respecte autant que possible la notion de voisinage : on passe alors d’un problème de segmentation sur deux dimensions à un problème de segmentation sur une seule dimension, comme précédemment. Les HMM permettent ainsi d’effectuer une segmentation en régions géographiques homogènes du point de vue des densités associées aux états. Pour ordonner les sites, nous utilisons une courbe de Peano qui est une courbe fractale et qui parcourt tous les points du plan en respectant la notion de voisinage spatial (cf. Fig. 8). La prise en compte du voisinage spatial dans cette approche n’est pas complète. Alors que deux points voisins sur la courbe sont voisins dans le plan, la réciproque n’est pas vraie. Dans les faits, comme le montrent des travaux récents [BEN 95], cette concession à la topologie n’affecte pas les résultats en segmentation d’images et en analyse de textures. La figure 9 montre une photo aérienne en 256 niveaux de gris ainsi que la classification obtenue à l’aide d’un HMM ergodique[4] de 5 états. La cartouche située en haut et à droite de la photo 9(b) montre en fonction du niveau de gris quelle est la transition de probabilité maximale. On ne montre pas les transitions entre états différents. On voit que les régions de texture homogènes sont associées au bouclage du processus stochastique sur chacun des 5 états. Dans une région uniforme, les niveaux de gris des pixels suivent la distribution associée à un état. Le choix du modèle ini-

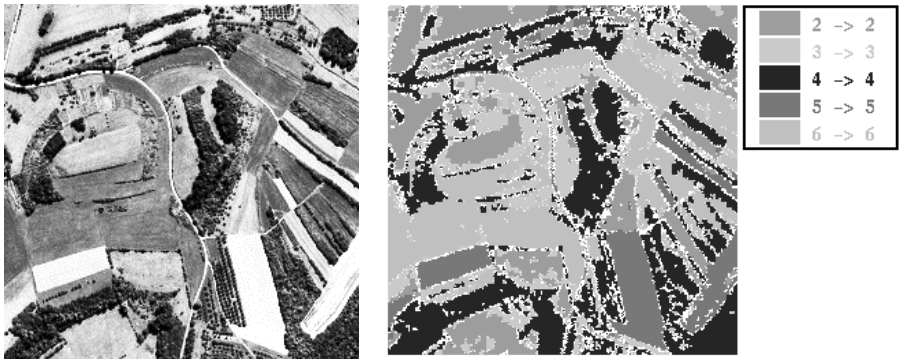
4. Tous les états sont inter-connectés. La matrice $\mathcal{A}$ ne contient pas de zéro.

tial qui conditionne le nombre de classes extraites est toujours un problème ouvert [BRE 01, PRE 98].



(a) image 2x2 (b) image 4x4 (c) image 6x6

Figure 8. *Ordonnancement des points du plan suivant une courbe fractale*



(a) original en 256 couleurs (b) segmentation par un HMM ergodique à 5 états

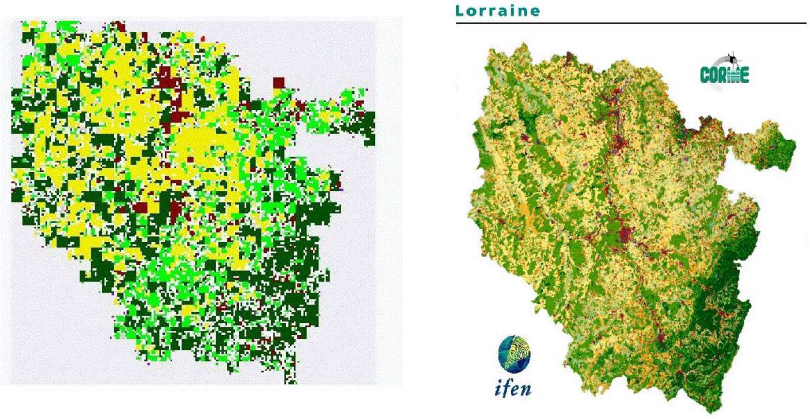
Figure 9. *Segmentation en 5 états d'une photo aérienne*

4.1. *Classification des données Ter-Uti*

Nous avons adapté la courbe de Peano pour pour parcourir l'ensemble des sites *Ter-Uti* en respectant les deux niveaux de résolution spatiale propres aux données *Ter-Uti* : un premier parcours permet d'introduire une relation d'ordre sur les 32×32 photos aériennes couvrant la Lorraine ou la région Midi-Pyrénées[5] ; un deuxième par-

5. Nous avons ajouté des photos "noires" pour obtenir un espace carré, nécessaire au parcours de Peano.

cours à l'intérieur des photos aériennes ordonne les sites *Ter-Uti* (cf. parcours donné Fig. 8(c)). Nous pouvons ainsi réaliser une segmentation spatiale des données : un point n'est plus classé en fonction de ses précédents temporels, mais en fonction de ses précédents spatiaux selon la courbe de Peano.



(a) Classification obtenue par un HMM2 (b) Corine Land Cover

Figure 10. *Classification obtenue par un HMM2 à 10 états après 10 itérations et comparaison avec la carte de la Lorraine établie dans le cadre du programme Corine Land Cover à partir de données satellitaires (IFEN, 1993)*

Avec un changement minime dans le programme d'apprentissage, nous avons traité l'ensemble des données, chaque année individuellement et toutes les années ensemble. Nous avons traité les données brutes (un site = une culture) et les données des successions (un site = une succession de cultures sur trois années). Ce sont ces dernières qui ont fourni les résultats les plus stables. Ce processus de classification est non supervisé ; l'utilisateur définit seulement le nombre initial d'états correspondant au nombre maximum de régions à découvrir. Nous obtenons des cartes (cf. Fig. 10(a) pour la Lorraine et Fig. 11 pour Midi-Pyrénés) où se distinguent plusieurs zones homogènes vis-à-vis de l'occupation du sol. Pour la Lorraine, nous distinguons cinq états principaux, les autres états provenant d'une division abusive par l'algorithme :

– un état à majorité de bâti (30 %), forêt, sols nus, zones humides, qui suit le cours des grandes vallées ;

– un état à majorité de forêt (98 %) qui englobe le massif vosgien et les forêts de la Meuse ;

– un état à majorité de prairies (30 %), forêt (20 %) et cultures fourragères (6 %) caractéristique des régions d'élevage ;

– un état à majorité de cultures (30 % pour blé, colza, et orge) et prairie (10 %) définissant les régions à dominante céréalière ;

– un état à majorité de prairies (68 %) et prés vergers (5 %) au fond des vallées vosgiennes et en pieds de côtes.

On retrouve globalement sur la figure 10(a) la localisation des occupations affichées sur la carte de la figure 10(b) telles qu'elles sont établies à partir de données satellitaires à une échelle 4 fois[6] plus fine (programme *Corin Land Cover*). La localisation des états est interprétable en fonction de la géologie et des caractéristiques techniques des exploitations. Ainsi en Lorraine, on retrouve les régions géologiques des Vosges gréseuses, des Vosges granitiques, du plateau lorrain (calcaire), des vallées de la Moselle et de la Meuse, des vallées vosgiennes, des plaines argileuses. On trouve aussi la distinction, sur le même terrain géologique (argiles de Keuper), entre les petites exploitations laitières du Châtenois et les grandes structures mixtes du Saulnois. Cette dernière distinction méritera d'être confortée par une étude sur les types d'exploitations agricoles en fonction des petites régions agricoles.

4.2. *Aide à l'interprétation d'images satellitaires en Midi-Pyrénées*

Les images satellitaires sont un moyen puissant pour connaître les occupations du sol dans une région et établir des prévisions en matière de production agricole ou de besoins en eau (cf. [CAS 98, GEN 01]). Un problème majeur tient pourtant au fait qu'à la date où les prévisions doivent être établies (en début de saison), les images satellitaires disponibles ne sont pas suffisantes pour cartographier avec certitudes les occupations du sol de l'année courante. La connaissance des successions de cultures pratiquées dans la région concernée permet alors de lever certaines ambiguïtés : connaissant l'occupation d'une parcelle l'année précédente, on réduit le champ des possibles pour l'année courante et une image de début de saison, qui permet de distinguer uniquement les cultures de printemps et cultures d'hiver, peut alors suffire pour conclure. Cette approche a été étudiée dans [LAR 00] avec un modèle qualitatif des successions de cultures.

Nous intervenons dans un projet qui a pour but de tester cette approche dans le cadre de la gestion de l'irrigation en Midi-Pyrénées. Il s'agit dans un premier temps de rassembler une carte d'occupation du sol à l'année $n-1$ et une image satellitaire de début de saison (avril) de l'année n. Dans un deuxième temps, il s'agit d'établir les probabilités de transition entre cultures à partir de données *Ter-Uti* sur la région. Le principe est alors de coupler les images et les probabilités de transition pour obtenir une estimation de la carte d'occupation du sol à l'année n. L'évaluation de l'estimation obtenue s'appuie sur une carte d'occupation du sol de l'année n établie *a posteriori*. Pour ce projet, les cartes d'occupation du sol peuvent être obtenues de différentes façons, par relevé de terrain ou par croisement de trois images satellitaires prises à différentes périodes.

6. Un point *Ter-Uti* représente 100 ha, contre 25 ha dans le programme *Corin Land Cover*.

En collaboration avec le centre INRA de Toulouse, nous avons donc effectué une classification automatique non supervisée de la région Midi-Pyrénées pour dégager trois régions homogènes caractérisées chacune par leur successions. Nous avons appliqué les mêmes traitements que pour la Lorraine tels qu'ils sont définis dans la section 4. Sur chaque région, nous avons ensuite procédé à une analyse plus fine afin d'étudier les évolutions temporelle des successions selon la démarche décrite en section 3. Le but est de construire le modèle 2 et son diagramme (cf. Fig. 4) sur chaque région afin d'avoir les valeurs des probabilités de transition entre cultures suivant l'endroit et l'année.

Sur la figure 11, du Nord au Sud, on distingue 3 régions homogènes :

années 90	
0.099	bois + bois + bois
0.089	maïs + maïs + maïs
0.082	prairies + prairies + prairies
0.034	blé + tournesol + blé
0.028	vignes + vignes + vignes
0.026	tournesol + blé + tournesol
0.018	jachères + jachères + jachères

années 2000	
0.105	maïs + maïs + maïs
0.099	bois + bois + bois
0.081	prairies + prairies + prairies
0.056	blé + tournesol + blé
0.037	tournesol + blé + tournesol
0.027	vignes + vignes + vignes
0.023	jachères + jachères + jachères

Figure 11. *Cartographie des successions culturales dans les départements du Gers et des Hautes-Pyrénées*

– au Nord, une première zone colorée en gris clair contient les cultures. Une modélisation de cette région par un modèle de Markov à trois états détecte trois périodes. On remarque dans les 2 tableaux de droite la progression de la rotation blé-tournesol et de la monoculture du maïs durant les 10 dernières années ;

– plus au Sud, une deuxième zone colorée en gris foncé dans laquelle la distribution des couverts reste stationnaire et est constituée de bois, de superficie en herbe et des premiers alpages ;

– enfin, une dernière zone colorée en noir contient les massifs montagneux.

On voit sur cet exemple l'intérêt d'utiliser le même modèle de segmentation dans les dimensions spatiales et temporelles des données. La segmentation spatiale permet de définir des régions homogènes vis-à-vis des successions sur une longue période puis une segmentation temporelle permet une étude plus fine région par région.

5. Discussions et conclusions

Nous avons développé une méthode de classification de données temporelles et spatiales qui est fondée sur le calcul d'une probabilité *a posteriori* obtenue à partir d'un modèle probabiliste – le HMM – qui prend en compte la variabilité temporelle et spatiale des données. La même mesure probabiliste sert à classer les données dans le temps et dans l'espace et donne une cohérence aux traitements.

Pour les agriculteurs, le choix des successions culturales est le révélateur *a posteriori* des ajustements qu'il réalisent entre leurs parcelles disponibles et l'importance surfacique des couverts végétaux qu'ils ont choisis. Une lecture de ces successions nous renseigne sur les difficultés et les simplifications auxquelles sont confrontés les agriculteurs dans leur ajustement "territoire – couverts végétaux".

D'autre part, les successions culturales sont des résultats d'organisation d'agriculteurs qui influencent fortement les effets de l'agriculture sur les ressources renouvelables :

– l'érosion (proportion des inter-cultures longues) ;

– la dégradation des qualités d'eaux souterraines ;

– les besoins en irrigation ;

– les monotonies paysagères.

Pour l'agronome, disposer des successions culturales à une échelle régionale permet de mieux choisir ses dispositifs. Qu'il s'agisse de prévisions, d'enquêtes ou de mise en place d'expérimentations, cet outil permet d'instruire le domaine de validité et d'extrapolation de ses études.

Enfin, il faut souligner que les successions culturales sont des objets de recherche centraux [SEB 88] révélateurs du métier d'agriculteur, mais qui ont donné lieu à peu de publications.

Dans le travail décrit, les allers et retours entre agronomes et informaticiens se sont traduits par l'élaboration de plusieurs modèles pour valider les hypothèses des agronomes. Nous nous sommes aidés de la durée des états comme révélatrice de l'instabilité des successions ainsi que du contenu de l'état de réserve qui capte toutes les exceptions et permet d'effectuer une segmentation progressive des données. À partir d'un modèle simple – le modèle 1 – donné dans les figures 1 et 2, nous avons élaboré les modèles 2 et 3 pour mesurer plus précisément des choix de successions. La figure 6 montre un système quasiment en équilibre. Les successions dominantes ont été trouvées, elle sont représentées par les états faiblement connectés entre eux. Ces traitements peuvent être effectués à différentes échelles, PRA ou région entière, pour peu que le nombre de points soit suffisants.

Les algorithmes développés sur les données lorraines ont pu être utilisés sur les données du bassin de la Seine ou en Midi-Pyrénées. Les agronomes se sont appropriés et ont amélioré les outils de modélisation et visualisation fondés sur les HMM mis à leur disposition [CAT 99, BOR 00]. Ils ont réussi à extraire des informations aux-

quelles les informaticiens n'avaient pas pensé *a priori*, telles des successions à long terme dépassant l'ordre du modèle. Le diagramme de la figure 12 est extrait d'une étude sur la qualité de l'eau dans les PRA de la région parisienne. Il a été annoté par et pour des agronomes.

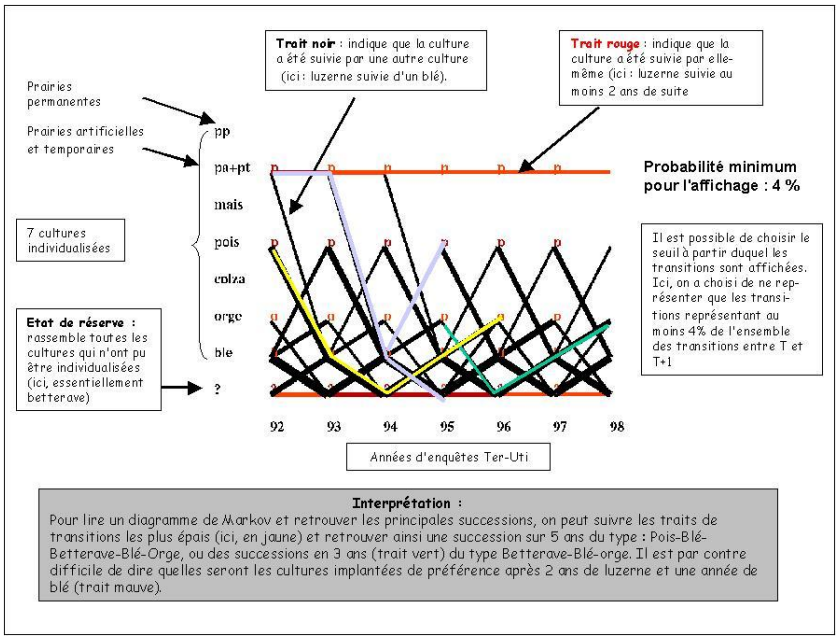


Figure 12. *Utilisation de l'outil de visualisation sur une PRA de la région parisienne*

La carte issue d'une segmentation 2D est un document porteur d'une information de synthèse appréciée des agronomes qui la comparent spontanément à d'autres cartes et ainsi la valident en partie. Ces résultats permettent de remettre à jour, confirmer ou infirmer les connaissances sur la région étudiée. En découle également la spécification d'une nouvelle classification dans laquelle nous devrons croiser différentes sources de données : les données *Ter-Uti*, les données issues d'enquêtes auprès des agriculteurs et relatives aux types d'exploitations agricoles mais aussi les cartes de différentes données pedo-morphologiques et les photos satellitaires. Ces connaissances peuvent entrer dans différents modèles [LeB 98, BEN 97] pour effectuer des simulations prospectives sur l'occupation agricole du sol et ses effets en matière d'eau et d'environnement.

Remerciements

Nous remercions les services régionaux des statistiques agricoles des DRAF Lorraine et Midi-Pyrénéespour l'accès aux données *Ter-Uti*.

6. Bibliographie

[BAK 74] BAKER J. K., « Stochastic Modeling for Automatic Speech Understanding », REDDY D., Ed., *Speech Recognition*, p. 521 – 542, Academic Press, New York, New-York, 1974.

[BEN 97] BENOÎT M., PAPY F., « Pratiques agricoles et qualité de l'eau sur le territoire alimentant un captage », *L'eau dans l'espace rural*, p. 323-338, INRA, 1997.

[BEN 01] BENOÎT M., LE BER F., MARI J.-F., « Recherche des successions de cultures et de leurs évolutions : analyse par HMM des données Ter-Uti en Lorraine », *Agreste Vision - La statistique agricole*, , n° 31, 2001, p. 23–30.

[BEN 95] BENMILOUD B., PIECZYNSKI W., « Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images », *Traitement du signal*, vol. 12, n° 5, 95, p. 433 – 454.

[BOR 00] BORNERAND C., « Dynamique des pratiques culturales dans le bassin de la Marne depuis les années 70 », Mémoire de DAA ENSAIA, septembre 2000, INRA SAD Mirecourt.

[BOY 90] BOYER A., MARTINO J. D., DIVOUX P., HATON J.-P., MARI J.-F., SMAILI K., « Statistical Methods in Multi-Speaker Automatic Speech Recognition », *Applied Stochastic Models and Data Analysis*, vol. 6, n° 3, 1990, p. 143–155, John Wiley and Sons, Ltd.

[BRE 01] BREHELIN L., « Modèles de Markov cachés et apprentissage par fusions d'états : algorithmes, applications, utilisations pour le test de circuits intégrés », PhD thesis, Univ. de Montpellier II, juin 2001.

[CAS 98] CASTERAD M., HERRERO J., « Irrivol : A method to estimate the yearly and monthly water applied in an irrigation district », *Water Resources Research*, vol. 34, 1998, p. 3045–3049.

[CAT 99] CATY M., « Évolution des pratiques agricoles et liens avec l'évolution de la qualité de l'eau dans le bassin de la Seine », Mémoire de fin d'étude ENGEES, 1999, INRA SAD Mirecourt.

[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., « Maximum-Likelihood From Incomplete Data Via The EM Algorithm », *Journal of Royal Statistic Society, Ser. B (methodological)*, vol. 39, 1977, p. 1 – 38.

[FOR 73] FORNEY G., « The Viterbi Algorithm », *IEEE Transactions*, vol. 61, 1973, p. 268–278.

[GEN 01] GENOVESE G., VIGNOLLES C., NEGRE T., PASSERA G., « A methodology for a combined use of normalised difference vegetation index and CORINE land cover data for crop yield monitoring and forecasting. A case study on Spain », *Agronomie*, vol. 21, 2001, p. 91–111.

[KIE 93] KIENAST F., « Analysis of historic landscape patterns with a geographical information system - a methodological outline », *Landscape Ecology*, vol. 8, n° 2, 1993, p. 103–118.

[LAR 00] LARGOUËT C., « Aide à l'interprétation d'une séquence d'images par la modélisation de l'évolution du système observé. Application à la reconnaissance de l'occupation du sol », Thèse de l'Université de Rennes I, novembre 2000.

[LeB 98] LE BER F., BENOÎT M., « Modelling the spatial organisation of land use in a farming territory. Example of a village in the Plateau Lorrain. », *Agronomie : Agriculture and*

Environment, vol. 18, 1998, p. 101-113.

[LED 92] LEDOUX M., THOMAS S., « De la photographie aérienne à la production de blé », *Agreste, la statistique agricole*, vol. 5, 1992.

[MAR 97] MARI J.-F., HATON J.-P., KRIOUILE A., « Automatic Word Recognition Based on Second-Order Hidden Markov Models », *IEEE Transactions on Speech and Audio Processing*, vol. 5, 1997, p. 22 – 25.

[MAR 00] MARI J.-F., LE BER F., BENOÎT M., « Fouille de données par modèles de Markov cachés », *Journées francophones d'ingénierie des connaissances (IC2000)*, mai 2000, p. 197–205.

[PRE 98] DU PREEZ J. A., « Efficient High-Order Hidden Markov Modelling », PhD thesis, University of Stellenbosh, 1998.

[RAB 95] RABINER L., JUANG B., *Fundamentals of Speech Recognition*, Prentice Hall, 1995.

[SEB 88] SEBILLOTTE M., « Les systèmes de culture », *Encyclopedia Universalis*, 1988.

[TOU 74] TOU J. T., GONZALES R., *Pattern Recognition Principles*, Addison-Wesley, 1974.